Diplôme d'Etudes Approfondies

Analyse et Modélisation des Systèmes Biologiques

Université Claude Bernard - Lyon I

Rapport bibliographique

Modèles hiérarchiques en épidémiologie vétérinaire

février 1994

Renaud LANCELOT
CIRAD-EMVT

Directeur de Recherches:	Pierre Charles LEFÈVRE, CIRAD-EMVT
Parrains Scientifiques:	Bernard FAYE, INRA
	Jean-Christophe THALABARD, URA CNRS 1454
	Jean-Pierre FLANDROIS, URA CNRS 243

Sommaire

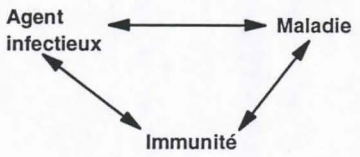
Introduction

Epidémiologie, systémique et hiérarchie des systèmes

L'épidémiologie peut être définie comme l'*étude de la distribution et des facteurs d'état ou d'événements de santé dans des populations déterminées, ainsi que l'application de cette étude à la maîtrise des problèmes de santé.*(1) **[1]**. Très souvent, les facteurs d'état ou d'événement de santé sont appelés **facteurs de risque**: ils sont associés à l'augmentation de la probabilité d'apparition ou de développement d'un phénomène pathologique. Cette définition générale s'applique sans difficulté à la pathologie animale.

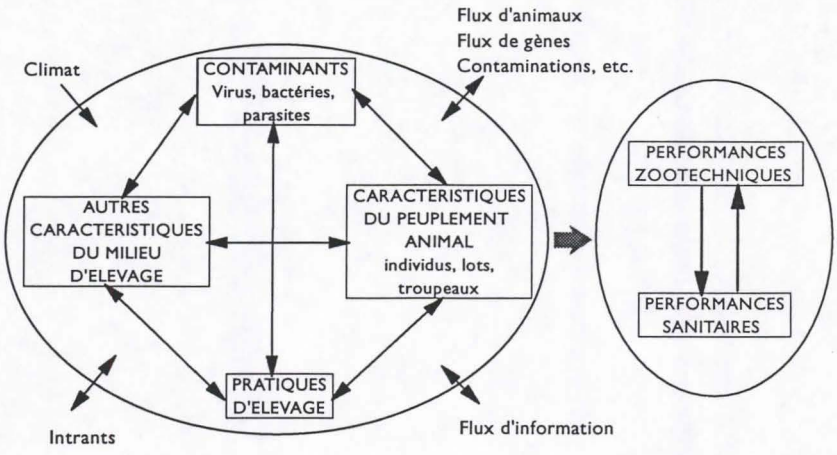
Beaucoup de pathologies observées en élevage sont multifactorielles: leur apparition ne peut être expliquée par l'action d'un unique agent pathogène, responsable de l'ensemble des symptômes observés. Ces maladies sortent du cadre du postulat de Koch qui énumère les conditions nécessaires pour établir une relation causale entre un agent pathogène supposé être responsable d'une maladie et cette maladie. Dans ces conditions, on ne peut se référer au modèle explicatif pasteurien, classiquement utilisé en épidémiologie vétérinaire (Fig. 1).

Fig. 1. Conception classique de la pathologie en épidémiologie vétérinaire



LANDAIS **[2]** lui substitue un schéma explicatif (Fig. 2) directement inspiré de la notion de système d'élevage, défini comme l'ensemble *des animaux et des techniques d'élevage pratiquées par l'éleveur dans le but d'obtenir des produits commercialisables et de maintenir ou d'accroître le potentiel de production* **[3]**.

Fig. 2. Modèle conceptuel de l'écopathologie (LANDAIS, 1991)



Ce schéma est celui de l'écopathologie, branche de l'épidémiologie vétérinaire ainsi définie par GANIERE *et al.* **[4]**: *appliquée aux productions animales,* [elle] *étudie, dans les élevages, l'ensemble des facteurs qui, en interaction dans l'environnement biologique, physique, humain et économique des animaux, est susceptible d'induire un état pathologique et/ou d'affecter leur productivité et la qualité des produits qui en dérivent.*

L'approche écopathologique est délibérément systémique, dans la mesure où elle considère la santé comme une des composantes du système d'élevage, que l'on ne peut appréhender isolément. En effet, selon MORLEY **[5]**, *puisque les différentes composantes*

(1) Dans ce document, les citations d'auteurs sont écrites en italique.

du système sont liées de manière interactive et interdépendante, l'examen d'une composante isolée peut conduire à des conclusions erronées car les réponses en retour à l'intérieur du système ne sont alors pas prises en compte.

Dans ce contexte, l'écopathologie est confrontée au principe de la hiérarchisation des systèmes biologiques. En effet, comme le notent MARTIN *et al.* **[6]**, une caractéristique majeure de ces systèmes *est de pouvoir être envisagés, conceptuellement et pratiquement, de manière verticale, en partant des atomes pour aller vers les cellules, les organes, les appareils, les individus et les populations...* Dans le cas présent, l'animal malade appartient à un lot (étable, poulailler...), lui même faisant partie d'une exploitation agricole. Des échelles supérieures peuvent être rencontrées: organisations d'éleveurs, zone écologique, etc. A chaque niveau de la hiérarchie correspondent des caractéristiques intrinsèques ou des pratiques d'élevage, individuelles ou collectives. L'épidémiologiste doit déterminer la part de la variabilité du phénomène à expliquer due:

☞ à l'individu (animal): sexe, âge, quantité d'aliment ingérée...

☞ au troupeau ou au lot de provenance: ambiance du bâtiment d'élevage, nature de l'alimentation, mesures de prophylaxie mises en oeuvre par l'éleveur...

☞ et à tout autre agrégat d'échelle supérieure qu'il serait pertinent de considérer.

La connaissance de la variabilité associée à chaque échelle de cette hiérarchie devrait permettre de déterminer les actions à mener en priorité dans le cadre d'un plan de prévention de la maladie étudiée.

Un autre élément structure les données: le temps chronologique. La considération de la chronologie des événements permet parfois de simplifier le modèle de départ: un facteur de risque survenant après l'apparition du phénomène étudié ne saurait être pris en compte dans l'analyse !

Restrictions apportées à l'étude

Nous limitons le sujet aux pathologies multifactorielles, se produisant dans des systèmes d'élevage à structure hiérarchique. Comme nous venons de l'indiquer, ces maladies sont du domaine de l'écopathologie. Nous n'envisageons que les *modèles d'association* **[7]**, *dont le but est de tenter d'établir l'étiologie d'une maladie en observant les associations entre sa survenue et la présence de divers facteurs de risque.*

Le rapport technique qui suivra ce travail concernera une enquête écopathologique de la mortalité des chèvres en zone péri-urbaine de N'Djaména (Tchad), effectuée pendant l'hiver 91-92. Dans la suite, cette enquête est nommée [Tchad]. Nous indiquons, le cas échéant, les particularités et les contraintes méthodologiques imposées par les milieux tropicaux difficiles **[8]**.

De plus, dans l'optique de cette application, nous nous intéressons essentiellement aux méthodes statistiques traitant des variables dichotomiques, bien adaptées à l'étude de la mortalité. Par ailleurs, nous n'abordons pas le problème du traitement des observa-

tions répétées dans le temps ou des suivis de cohorte sur de longues périodes.

Plan du rapport

En premier lieu, nous examinons la place et les conséquences de la hiérarchie dans la conception d'une enquête épidémiologique. Nous commençons par la définition de la problématique de l'étude: choix d'une échelle d'observation et approche statistique. Nous envisageons ensuite la manière d'écrire le pré-modèle conceptuel d'analyse. Nous montrons que ce schéma est bien adapté pour rendre compte de la hiérarchie des données imposée par le temps et permet de simplifier le réseau des interactions statistiques théoriquement envisageables.

Puis nous décrivons les méthodes statistiques habituellement utilisées pour mesurer le risque associé au pré-modèle conceptuel d'analyse: régressions linéaire et logistique. Nous en indiquons les limites dans le cadre d'effets de groupe ainsi que les aménagements employés pour y remédier. Nous soulignons que ces modèles sont mal adaptés à la prise en compte des hiérarchies autres que celles induites par des considérations de logique ou de chronologie.

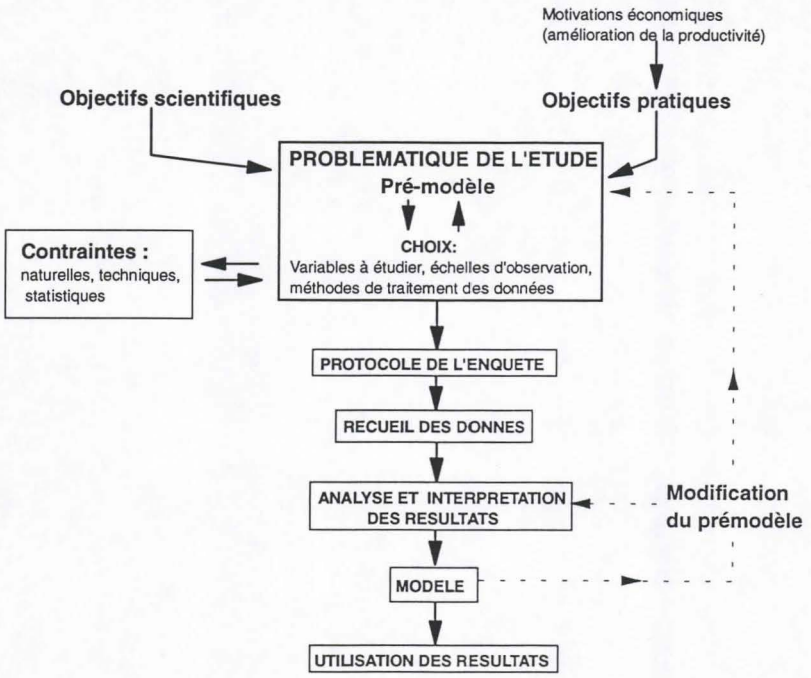
Dans une troisième partie, nous étudions des modèles hiérarchiques plus généraux, mieux adaptés à traiter les structures hiérarchiques rencontrées dans les systèmes d'élevage, mais dont l'usage est encore très limité en épidémiologie vétérinaire. Nous présentons un modèle d'analyse hiérarchique bayésienne et deux modèles de régression hiérarchique: un modèle linéaire et un modèle non linéaire.

En conclusion, nous faisons le bilan des méthodes actuellement utilisées pour traiter les systèmes hiérarchiques, de celles qu'il serait souhaitable d'employer et de celles qui sont effectivement à la disposition des praticiens de l'analyse des données. Enfin, nous indiquons ce que nous souhaitons réaliser dans le cadre du mémoire pratique.

1. La hiérarchie dans la construction du pré-modèle conceptuel d'analyse

FRONTIER **[9]** rappelle qu'*un modèle, en tant qu'outil de recherche, doit correspondre à un objectif.* Le modèle est l'aboutissement d'un processus dont il est indissociable, résumé dans le schéma ci-contre.

Fig. 3. Processus d'élaboration d'un modèle (d'après FRONTIER)



Une étape essentielle de la construction du modèle est la définition de la problématique de l'étude et l'élaboration d'un schéma: le **pré-modèle conceptuel d'analyse**, qui visualise les hypothèses de facteurs de risque.

Outre l'organisation du système d'élevage que nous avons évoquée ci-dessus, d'autres éléments structurent les données et doivent être pris en compte dès cette étape. Il s'agit de l'ordre chronologique dans lequel se produisent les événements étudiés (un effet ne peut pas précéder une cause), ainsi que de leur ordre logique.

1.1. La hiérarchie et la problématique de l'étude

La question posée peut être envisagée selon 2 points de vue différents: opérationnel et biologique.

☞ **Point de vue opérationnel:** en épidémiologie vétérinaire en général, et dans les pays peu développés en particulier, l'objectif opérationnel est d'identifier les facteurs de risque d'une pathologie dominante, afin de proposer un programme de lutte basé sur des actions techniques simples et peu coûteuses. Il est vraisemblable que les améliorations les plus efficaces seront celles s'adressant d'emblée à une population ou une sous-population animale plutôt qu'à des individus: mesures de prophylaxie médicale collective (vermifuge, vaccin), modification de pratiques d'élevage (ventilation des bâtiments, renouvellement de la litière...). L'adoption de ce point de vue conduirait donc à considérer cette population comme objet d'étude, d'analyse et d'action.

☞ **Point de vue biologique:** à l'inverse, en termes biologiques, un phénomène pathologique se réalise, ou ne se réalise pas, à l'échelle de l'individu (et non pas à celle d'une population ou d'une sous-population). Il est donc logique que les hypothèses de facteurs de risque se rapportent directement à l'individu, et que le modèle statistique soit élaboré à ce niveau.

Cependant, dans la réalité, le point de vue opérationnel dicte le choix d'une échelle supérieure. FRONTIER **[8, 9]** indique en effet que *lorsque l'objectif de l'étude est pratique, l'échelle d'observation est imposée par le type de système que l'on a décidé de gérer.*

1.1.1. Faut-il étudier spécifiquement chaque niveau hiérarchique ?

Le choix du troupeau comme échelle d'observation peut entraîner des difficultés statistiques. Ainsi, quand le modèle est élaboré à l'échelle de l'individu (point de vue biologique) alors que l'échantillonnage initial a porté sur des troupeaux (point de vue opérationnel), voire des villages, il est vraisemblable que l'hypothèse d'indépendance des observations n'est pas respectée. Or, cette hypothèse est fondamentale dans de nombreuses méthodes statistiques (notamment les techniques de régression). CORNFIELD **[10]** démontre que *la pratique qui consiste à échantillonner des groupes puis à effectuer une analyse à l'échelle de l'individu sans précaution, donne des résultats décevants et devrait être proscrite.*

Pour pallier ces difficultés, BENDIXEN **[11]** préconise la réalisation d'enquêtes cas-témoins différentes pour chaque niveau de la hiérarchie. Pour la mise en évidence des facteurs de risques individuels, il recommande d'échantillonner le témoin dans le même troupeau que le malade. Pour ceux relevant du troupeau, il suggère de stratifier ces derniers selon l'incidence de la pathologie étudiée et de comparer les strates extrêmes. Les troupeaux à forte incidence sont les cas, et les troupeaux à faible incidence sont les témoins. Pour standardiser les effets individuels à l'intérieur de chaque troupeau, BENDIXEN conseille de travailler sur des "sous-troupeaux" homogènes vis-à-vis du (des) facteur(s) de risque individuel(s) considéré(s).

Cependant, cette façon de procéder peut se révéler impraticable. Si plusieurs facteurs de risque sont importants à l'échelle individuelle: les strates contiennent un très faible nombre de sujets. De plus, elle ne répond pas à nos préoccupations: elle ne permet pas de déterminer la part de la variabilité imputable à chaque échelle, ni de comparer entre elles les covariables de niveaux différents. Par ailleurs, elle impose des conditions d'échantillonnage incompatibles avec les conditions de l'Afrique **[8]**.

La formulation statistique du modèle doit refléter la hiérarchie du système, c'est-à-dire traiter différemment les variables explicatives selon l'échelle à laquelle elles se réfèrent.

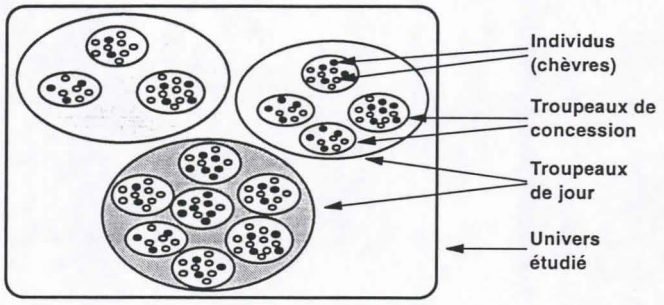
1.1.2. Choix de l'échelle d'observation et stratégie d'échantillonnage. Contraintes en milieu difficile

Dans l'enquête [Tchad], le système étudié est constitué de 3 sous-systèmes emboîtés (Fig. 4):

- ☞ l'individu (chèvre),
- ☞ le troupeau de concession (groupe d'animaux passant la nuit près de l'habitation d'un même éleveur),
- ☞ le troupeau de jour (rassemblement de plusieurs troupeaux de concession conduits au pâturage par un même berger).

L'échelle d'observation la plus pertinente en termes opérationnels est celle du troupeau de concession. Concrètement, cela signifie que la population de référence devrait être l'ensemble des troupeaux de concession de la région enquêtée. L'échantillon étudié serait alors obtenu par sondage aléatoire parmi cette population. Ceci pose peu de problèmes dans les pays développés, où tous les élevages sont répertoriés. Il n'en va pas de même dans les pays en développement: les infrastructures routières sont faibles (problèmes de déplacement) et les éleveurs ne sont pas recensés (pas de base de sondage). Pour des raisons pratiques, l'échelle retenue pour l'échantillonnage est le village, indépendamment des considérations sur le système d'élevage.

Fig. 4. Schéma d'un système d'élevage: enquête [Tchad]



Cela revient à faire un sondage en grappe, et rejoint ainsi les recommandations de SCHERRER **[12]**: *L'échantillonnage par degré s'impose lorsque les objectifs de l'étude visent l'estimation de paramètres aux différents niveaux d'unités...* [Il] *est fortement recommandé lorsqu'il est plus commode de prélever de façon aléatoire des grappes (voires des supergrappes) puis des éléments au sein de ces grappes que des éléments directement.*

Cette solution a été retenue dans l'enquête [Tchad]: l'échantillon de troupeaux de concession (échelle d'observation) a été obtenu par un sondage à 2 degrés (sondage par grappes). Le village représente le 1^{er} degré, et le troupeau de concession représente le $2^{ème}$ degré. Dans chaque troupeau, l'échantillonnage a été exhaustif: tous les individus ont été suivis. Dans d'autres études, on aurait pu envisager un échantillonnage des individus, introduisant ainsi un $3^{ème}$ degré dans le sondage.

1.2. Ecriture du pré-modèle conceptuel d'analyse

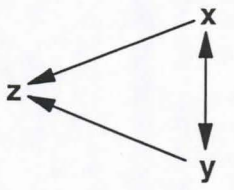
Cette phase est en fait contemporaine des étapes précédemment décrites. Il s'agit de résumer visuellement, sous forme d'un schéma, toutes les hypothèses des interactions étiologiques entre variable à expliquer et variables explicatives. Les écologistes comme GIGON **[13]** qualifient parfois ce schéma de "modèle hiérarchique de la toile d'araignée cybernétique des relations étiologiques et associées" (*model of the hierarchic cybernetic web of causal and correlative factors).*

En épidémiologie, un vocabulaire plus concis a été adopté: *path diagram* en anglais, que JENICEK *et al.* **[14]** ont traduit "schéma des interactions étiologiques". En général, les auteurs se servent de ces schémas dans le cadre de l'analyse des interactions étiologiques (*path analysis* en anglais), dont la théorie a été formulée par WRIGHT, un généticien, dès 1921 **[15]**. Il fallut attendre 1972 pour que cette technique soit introduite en épidémiologie, par GOLDSMITH et BERGLUND **[16]** qui ont d'emblée remarqué son intérêt pour l'étude de systèmes complexes et/ou hiérarchisés: *l'analyse des interactions étiologiques nous aide à déterminer les relations étiologiques vraisemblables entre une série de variables indépendantes et une ou plusieurs variables de santé dépendantes... On établit tout d'abord la liste des variables supposées importantes et* ***on les structure d'une certaine manière.***

1.2.1. Chronologie et logique du déroulement des événements

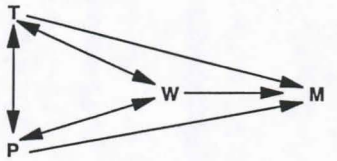
En 1977, GOLDSMITH **[17]** formalise les règles d'écriture de ces schémas en épidémiologie et justifie l'emploi de la méthode. Avec la régression multiple pas-à-pas (très employée en épidémiologie), on suppose implicitement que 2 variables explicatives **x** et **y** ont des relations symétrique entre elles et asymétrique avec la variable à expliquer **z** (Fig. 5). L'algorithme choisit entre **x** et **y** la variable ayant la plus grande variance dans la covariance commune de **x** et **y** avec **z**, en l'absence de considérations logiques ou biologiques.

Fig. 5. Interrelations entre variables: régression pas-à-pas



GOLDSMITH s'appuie sur un exemple fictif de l'étude de la mortalité quotidienne **M** dans les grandes villes, en fonction de la température **T**, la pollution **P** et les conditions météorologiques **W**. Le schéma complet des interactions, implicitement accepté quand on effectue une régression multiple pas-à-pas, est indiqué ci-contre.

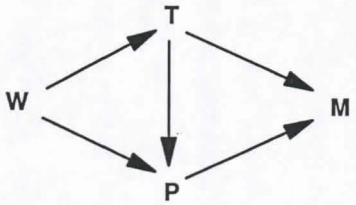
Fig. 6. Schéma des interactions implicites



La chronologie des événements observés au cours de l'enquête peut entraîner des simplifications: si de forts niveaux de pollution étaient observés avant la survenue d'un changement de température, la relation **T → P** n'aurait pas de sens.

Si les observations ou des connaissances antérieures permettaient de penser que les conditions météorologiques provoquaient à la fois un changement de température et une forte pollution, on proposerait comme schéma explicatif de la mortalité le diagramme ci-contre: toutes les relations sont unidirectionnelles et certaines relations théoriquement possibles (comme **W → M**) ne sont pas retenues.

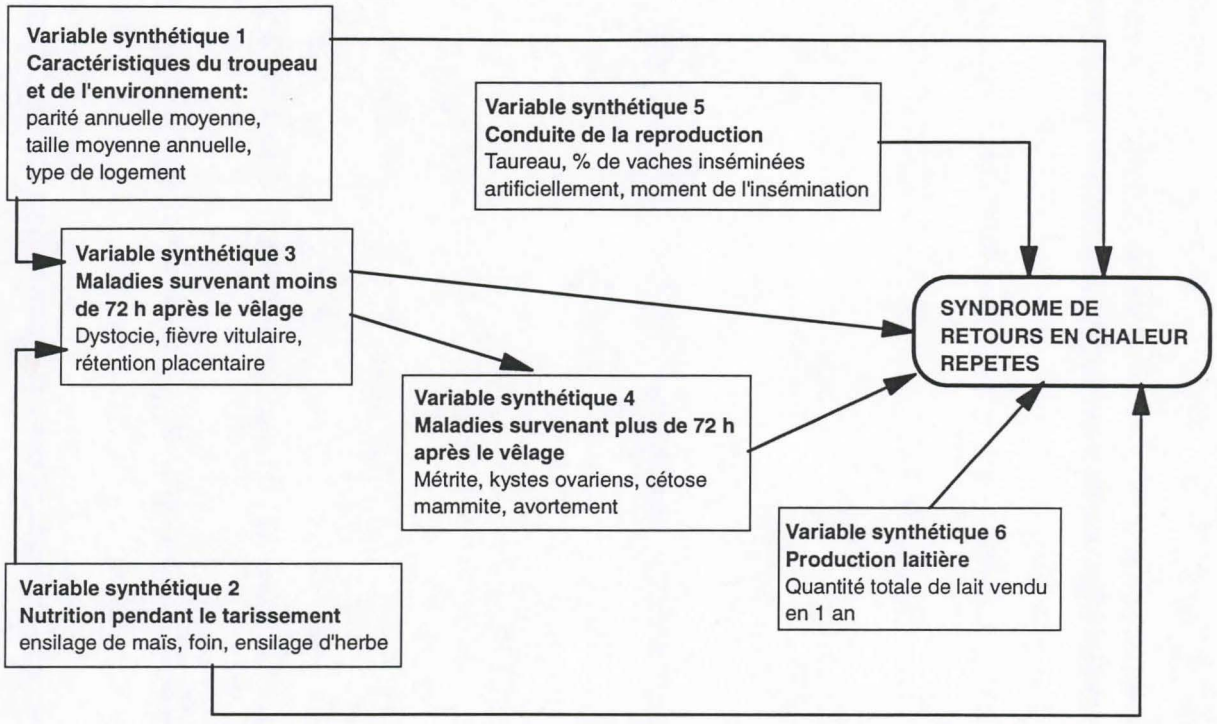
Fig. 7. Schéma des interactions plausibles



Selon une convention consacrée par l'usage, la variable dépendante est placée à droite, les variables explicatives les plus à gauche étant celles dont l'action est la plus ancienne. L'axe **gauche → droite** représente donc la structure conférée par le temps. L'ordre chronologique de survenue des événements est d'ailleurs l'élément hiérarchique le plus souvent pris en compte par les auteurs qui utilisent l'analyse des interactions étiologiques. Ils ont parfois du mal à rendre compte des autres hiérarchies auxquelles ils sont confrontés.

On peut le constater sur un exemple (Fig. 8) tiré d'un article de LAFI et KANEENE **[18]**, consacré à une étude du syndrome de retours répétés en chaleurs dans des élevages bovins laitiers, aux Etats-Unis. La définition du syndrome est pragmatique: l'amalgame est fait entre l'échec de l'insémination et les avortements précoces. Les variables explicatives relevées sont des facteurs de risque bien connus, relatifs à l'environnement, la nutrition, la pathologie, les pratiques d'élevage et la production laitière. Les objectifs de l'analyse sont de quantifier les relations entre le syndrome et ces groupes de facteurs de risque, et de proposer un modèle prédictif de l'incidence de ce syndrome par élevage. L'échelle privilégiée de l'analyse est donc le troupeau.

Fig. 8. Diagramme des interactions: enquête retours répétés en chaleurs (LAFI et KANEENE)

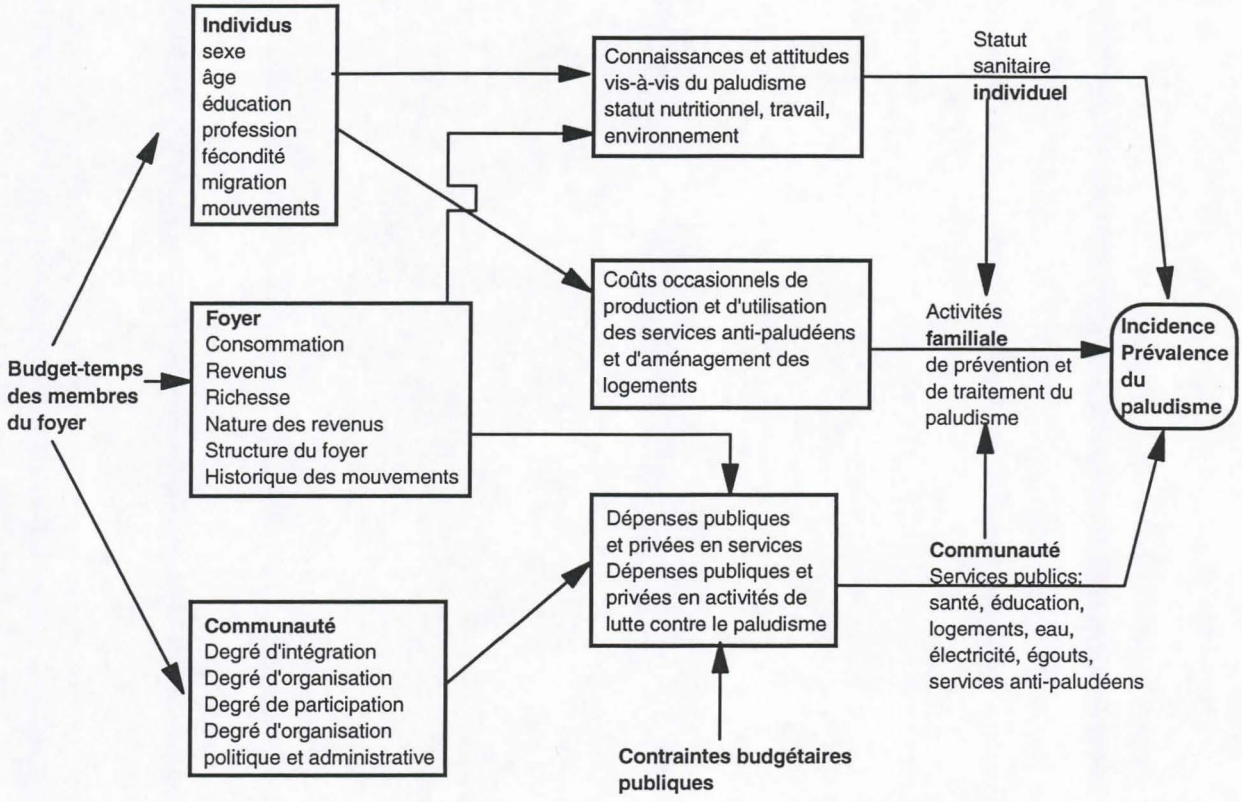


Des variables synthétiques sont construites à partir des facteurs de risque, dans le but de réduire le nombre des interactions à envisager et de faciliter l'interprétation des résultats. Trois variables synthétiques sont relatives aux individus (3, 4 et 6), alors que les autres (1, 2 et 5) décrivent l'environnement, le troupeau et les pratiques d'élevage. La hiérarchie n'est pas explicitée: les auteurs parlent uniquement de variables endogènes (individus) et exogènes (troupeau, conduite, milieu). La traduction graphique est confuse: dans le diagramme, les variables synthétiques décrivant les individus sont mélangées avec celles décrivant les pratiques d'élevage. De plus, la variable synthétique 1 (caractéristiques du troupeau et environnement) est hétérogène: on y trouve à la fois des descripteurs de la structure du troupeau et de son environnement (logement), qui ont peu de raisons d'être liés. Dans ce papier, la seule hiérarchie considérée comme importante est celle induite par la chronologie des événements. L'organisation du système étudié en troupeaux et individus dans les troupeaux est négligée.

1.2.2. Figuration des autres hiérarchies structurant le système

Dans un article consacré à la mise au point d'un modèle d'étude des facteurs socio-économiques favorisant le paludisme en Colombie **[19]**, BANGUERO propose un pré-modèle (Fig. 9) pour un système composé de 3 sous-systèmes emboîtés: individu, foyer familial et communauté. L'échelle privilégiée est celle du foyer. Dans ce modèle, la hiérarchie est figurée sur l'axe vertical (**Communauté → Foyer → Individus**, de bas en haut), alors que l'axe horizontal, plutôt que de représenter la chronologie des événements, détaille la stratégie de l'analyse: synthèse et agrégation des informations les plus à gauche (informations de base) pour simplifier l'analyse finale. Cela correspond aux variables synthétiques de LAFI et KANEENE, mais l'organisation hiérarchique socio-familiale est respectée. En revanche, la chronologie des événements n'est pas prise en compte.

Fig. 9. Modèle d'étude des facteurs favorisant le paludisme en Colombie (d'après BANGUERO)



2. Modèles statistiques pour l'étude des hiérarchies logique et temporelle

L'ordre chronologique et/ou logique de survenue des éléments est la hiérarchie la plus simple et la plus générale que l'on puisse considérer. Cette hiérarchie est prise en compte dans l'**analyse des interactions étiologiques**. Historiquement, cette technique s'est d'abord appuyée sur la régression linéaire multiple.

2.1. Régression linéaire

WRIGHT travaillait sur des variables quantitatives. Il utilisait la régression linéaire et estimait les paramètres par la méthode des moindres carrés. Dans ce cas, l'analyse des interactions étiologiques conduit à calculer des coefficients d'interactions $\mathbf{P_{ij}}$. Ce sont des nombres mesurant la proportion de la variabilité de la variable dépendante **i directement** expliquée par la variable indépendante **j**.

Concrètement, on peut schématiser l'analyse de la sorte (voir aussi Fig. 10), d'après GOLDSMITH **[17]**:

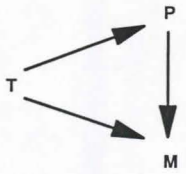
☞ On effectue tout d'abord plusieurs régressions partielles. Pour chacune de ces régressions, on prend comme variable dépendante une variable du pré-modèle à laquelle aboutissent directement les pistes d'autres variables. Les covariables de chaque régression partielle sont toutes les variables dont les pistes mènent directement à la covariable considérée. On obtient à l'issue de cette première étape un ensemble de coefficients de régression.

☞ On pose ensuite que les coefficients de régression sont des combinaisons linéaires des P_{ij}, en prenant comme règle que les P_{ij} des interactions en parallèle s'additionnent, alors que les P_{ij} des interactions en série se multiplient. Les P_{ij} sont calculés par la résolution du système d'équations linéaires.

Fig. 10. Principe de l'analyse des interactions étiologiques

Exemple:

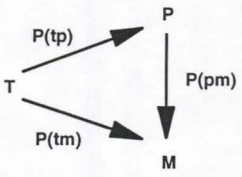
On considère le système des 3 variables T, P et M (schéma ci-contre).
On peut envisager 2 régressions partielles:
- **variable dépendante M en fonction des régresseurs P et T**
- **variable dépendante P en fonction du régresseur T.**



La réalisation de la première régression amène le calcul de 2 coefficients de régression: R(tm) et R(pm).
La réalisation de la seconde régression amène le calcul d'un coefficient de régression: R(tp)

On pose le système d'équations linéaires suivant:
- **P(mp) + P(mt) * P(pt) = R(tm)**
- **P(mp) + P(mt) * P(tp) = R(pm)**
- **P(tp) = R(tm)**

La résolution de ce système d'équations permet d'obtenir les 3 coefficients d'interaction étiologique recherchés: P(tp), P(pm) et P(tm) (schéma ci-contre).



☞ On peut alors décomposer l'effet total d'une variable explicative en effets directs et indirects. Cela permet de simplifier le schéma des interactions étiologiques en éliminant les pistes dont les effets directs associés sont faibles par rapport aux effets indirects. Le modèle statistique final est différent du pré-modèle. Il est d'ailleurs tout-à-fait possible de réitérer les calculs en substituant au pré-modèle initial ce modèle final.

Outre la simplification du modèle de départ, cette analyse permet donc d'obtenir un classement des variables explicatives selon l'intensité de leur effet total (décomposé en effets direct et indirects), répondant ainsi en partie à l'objectif opérationnel de l'étude.

Toutefois, il faut souligner, ainsi que le font ALWIN et HAUSER **[20]**, que tout raisonnement étiologique doit être mené en se référant au modèle étudié: il subsiste toujours la possibilité qu'une variable sous-jacente, non prise en considération dans le modèle, explique en fait tout ou partie de l'effet direct d'une variable du modèle.

De plus, comme toute technique reposant sur la régression linéaire, l'analyse des interactions étiologiques est basée sur l'indépendance des observations. Nous avons vu que cette condition n'est pas vérifiée dans les systèmes hiérarchisés, si l'échelle d'observation privilégiée est différente de l'individu. Divers travaux théoriques ont été entrepris pour corriger cette conséquence de la hiérarchie.

DONNER **[21]** traite ainsi, dans le cadre du modèle linéaire, des *méthodes appropriées d'analyse de données provenant d'un échantillonnage de groupes de sujets, ou de manière similaire, d'études non expérimentales dans lesquelles le principal traitement correspond à une caractéristique s'appliquant à l'échelle du groupe*: **G** modalités d'un traitement α sont étudiées dans le cadre de l'analyse d'une variable à expliquer **Y**.

Le but est d'ajuster le traitement α sur les X_u covariables individuelles, afin de pouvoir tester la signification de l'effet de ce traitement. DONNER propose un modèle à effets mixtes (effets fixes et effet aléatoire), qu'il écrit ainsi pour le $k^{ème}$ individu du $j^{ème}$ groupe de la $i^{ème}$ valeur de G:

$$Y_{ijk} = \mu + \alpha_i + \beta_1 X_{1ijk} + \beta_2 X_{2ijk} + \ldots + \beta_h X_{hijk} + V_{ij} + e_{ijk}$$

où: μ constante globale,
α_i effet de la i^{eme} modalité du traitement, en supposant $\Sigma \alpha_i = 0$,
β_u coefficient de régression partiel de Y sur X_u,
V_{ij} effet aléatoire du groupe, supposé ~ $N(0, \sigma_V^2)$,
e_{ijk} erreur individuelle, supposé ~ $N(0, \sigma_e^2)$.

Le coefficient de corrélation intragroupe est:

$$\rho = \frac{\sigma_V^2}{\sigma_V^2 + \sigma_e^2}$$

DONNER pose $W_{ijk} = V_{ij} + e_{ijk}$, terme d'erreur global, d'espérance nulle et de variance $\sigma_W^2 = \sigma_V^2 + \sigma_e^2$. Le modèle initial s'écrit alors:

$$Y_{ijk} = \mu + \alpha_i + \beta_1 X_{1ijk} + \beta_2 X_{2ijk} + \ldots + \beta_h X_{hijk} + W_{ijk}$$

En écrivant ce modèle, l'hypothèse nulle H_0 que l'on cherche à tester est:

$$\alpha_i = 0, \text{ avec } i = 1, 2, \ldots, G.$$

☞ si $\rho = 0$ (pas de corrélation intragroupe), le modèle se ramène à un modèle habituel d'analyse de la covariance;

☞ dans le cas général où $\rho > 0$, les techniques usuelles d'estimation des paramètres par la méthode des moindres carrés ordinaires conduisent à une surestimation de la signification statistique: la variance de ces paramètres est sous-estimée.

La solution consiste à utiliser pour l'estimation des paramètres non pas la méthode des moindres carrés ordinaires mais celle des moindres carrés généralisés. La seconde méthode permet en effet de prendre en compte la corrélation intragroupe, à la différence de la première. Divers logiciels peuvent effectuer ces calculs (SuperCarp©, GLIM©,...).

2.2. Régression logistique

Nous avons limité notre exposé à des variables dépendantes dichotomiques (malade/non malade, survie/mort). Le recours aux techniques présentées dans le paragraphe précédent n'est pas possible, car l'hypothèse de normalité des termes d'erreur n'est pas vérifiée: leur distribution est binômiale **[22]**. L'utilisation du modèle logistique permet d'éviter ce problème.

Rappels sur le modèle logistique

Le modèle logistique est une équation de régression entre une variable

dépendante qualitative à 2 classes et **n** variables explicatives $\mathbf{V_i}$ (i = 1,... n) qui peuvent être qualitatives ou quantitatives (BOUYER, **[23]** et **[24]**). Le modèle exprime la probabilité d'être malade connaissant les valeurs des $\mathbf{V_i}$, par une relation de la forme:

$$Pr(M = 1 \mid V_1, V_2, \ldots, V_n) = P = \frac{1}{1 + e^{-(\alpha + \sum_{I=1}^{n} \beta_I V_I)}}$$

Si l'on pose:

$$Logit\ P = Ln\left(\frac{P}{1-P}\right)$$

alors on peut écrire:

$$Logit\ P = \alpha + \sum_{i=1}^{P} \beta_i V_i$$

La principale raison du succès du modèle logistique en épidémiologie est que les paramètres $\boldsymbol{\beta_i}$ représentent, sous certaines conditions(2), les logarithmes des odds ratios (OR) associés aux covariables $\mathbf{V_i}$, ajustés sur l'ensemble des autres covariables. Leur interprétation biologique est donc aisée.

Utilisation dans l'analyse des interactions étiologiques

Récemment, CURTIS *et al.* (**[25]**, **[26]**) ont introduit ce modèle dans l'analyse des interactions étiologiques en épidémiologie vétérinaire. Les OR étant des mesures de risque multiplicatives, l'amplitude des associations directes et indirectes peut être mesurée en assimilant les OR ou les risques relatifs (RR) à des coefficients d'interactions étiologiques. Les associations directes sont représentées par les coefficients d'interaction (OR ou RR); les associations indirectes sont obtenues en multipliant les coefficients d'interaction le long des pistes car, avec la régression logistique, ces coefficients sont ajustés sur toutes les autres variables.

De manière analogue à ce que l'on a vu à la fin du paragraphe précédent, se pose le problème de la non-indépendance des observations dans les systèmes hiérarchisés. CURTIS *et al.* **[27]** soulignent que *l'utilisation de la régression logistique multiple nécessite la condition que les animaux possédant le même vecteur de covariables soient distribués indépendamment et de manière homogène. Le non respect de ces conditions modifie le résultat de l'analyse.*

Diverses variantes du modèle logistique ont été proposées pour pallier cette difficulté. Dans l'article que nous venons de citer, CURTIS *et al.* en ont comparé quelques unes. L'approche la plus satisfaisante semble être l'introduction dans le modèle d'une terme de **variance extra-binômiale (VEB)**.

---

(2) En particulier: les covariables doivent être qualitatives, et si elles présentent plus de 2 modalités, le codage de ces modalités doit obéir à certaines règles. Si les covariables sont quantitatives, les β_i ne peuvent pas être assimilés au logarithme d'un odds ratio.

L'apparition d'une VEB se produit quand les groupes sont hétérogènes vis-à-vis du risque de développer la maladie, c'est-à-dire quand les animaux appartenant à différents groupes caractérisés par le même vecteur de covariables, n'ont pas la même probabilité de développer la maladie. En d'autres termes, une ou plusieurs covariables n'ont pas été prises en compte dans le pré-modèle: il y a des facteurs de confusion sous-jacents. C'est bien ce qui se produit quand on effectue des régressions logistiques partielles à l'échelle de l'individu alors qu'il existe un effet troupeau. La VEB peut être d'origine très variable (écologique, génétique,...). La présence d'une VEB dans des données analysées par une régression logistique ordinaire entraîne une surestimation de la signification statistique des paramètres: la valeur moyenne du paramètre est peu modifiée, mais l'écart-type est sous-estimé.

Dans un article de synthèse sur ce problème, WILLIAMS **[28]** présente un modèle logistique permettant la prise en compte d'une VEB: on suppose que la $\mathbf{i^{ème}}$ réponse ($1 \leq i \leq n$) est un comptage de $\mathbf{R_i}$ succès et $\mathbf{m_i - R_i}$ échecs, et que cette réponse est associée aux valeurs ($\mathbf{x_{i1}, x_{i2}, ..., x_{ip}}$) de p variables explicatives. Quand on utilise un modèle logistique classique, on suppose que les $\mathbf{R_i}$ sont indépendamment distribués selon une loi binômiale de paramètres $\mathbf{m_i}$ et $\boldsymbol{\theta_i}$, avec:

$$\theta_i = \frac{1}{1 + e^{-\lambda_i}}, \qquad et \qquad \lambda_i = \sum_{s=1}^{p} x_{is}\beta_s$$

Afin d'introduire une VEB, on ajoute au modèle une variable P_i, distribuée sur l'intervalle [0, 1], d'espérance $\mathbf{E(P_i) = \theta_i}$ et de variance $\mathbf{Var(P_i) = \phi\, \theta_i\, (1 - \theta_i)}$. On suppose de plus que:

☞ les différentes P_i sont indépendantes
☞ P_i étant fixée et de valeur p_i, R_i suit une loi binômiale de paramètres m_i et p_i.

Il suit que:

$$E(R_i) = \theta_i, \; et \; Var(R_i) = \frac{m_i\theta_i(1 - \theta_i)}{1 + \phi(m_i - 1)}$$

La non indépendance entre observations s'exprime par un coefficient de corrélation intragroupe $\phi > 0$.

WILLIAMS indique également la différence entre les méthodes de calcul des paramètres $\boldsymbol{\beta_s}$, selon que les m_i sont identiques ou non, et selon que ϕ est connu ou non. Les calculs sont complexes et font intervenir le maximum de vraisemblance ou de quasi-vraisemblance. Ils sont programmés dans plusieurs logiciels commerciaux (GLIM©, EGRET©,...).

2.3. Limites de l'analyse des interactions étiologiques

L'analyse des interactions étiologiques a l'avantage d'aboutir à un modèle

statistique parfaitement cohérent avec le pré-modèle. Comme le schéma de FRONTIER l'indique (Fig. 3, ?), le processus d'élaboration du modèle peut d'ailleurs donner lieu à une ou plusieurs itérations avant d'aboutir à un résultat considéré comme satisfaisant par le chercheur. Les hiérarchies introduites par la chronologie des événements et la logique sont donc bien prises en compte.

Cependant, ces hiérarchies ne sont pas les seules à structurer les systèmes d'élevage. En particulier, elles ne décrivent les systèmes emboîtés (l'animal est dans un troupeau qui appartient à une exploitation agricole, par exemple). Les techniques telles que l'introduction dans le modèle d'un terme de VEB permettent de corriger la non-indépendance des observations individuelles, mais n'autorisent pas l'explication de la variance due à l'effet troupeau. Il y a alors une mauvaise adéquation entre la question posée (dans ce rapport) et les méthodes que nous venons de présenter.

Cette situation pose problème aux auteurs qui essaient de rendre compte d'une telle hiérarchie en employant l'analyse des interactions étiologiques. Dans certains articles, la question est éludée: les variables de différentes échelles d'observation sont considérées sur le même plan dans le modèle statistique. Au mieux, comme dans 2 publications de CURTIS *et al.* concernant les facteurs de risque de la morbidité et de la mortalité de veaux Holstein aux Etats-Unis, un modèle est-il élaboré (pré-modèle puis modèle statistique) pour chaque échelle d'observation: individu et troupeau (**[26]**, **[29]**). Cependant, les auteurs placent les mêmes covariables et les mêmes variables à expliquer dans les 2 modèles: les résultats sont difficiles à interpréter car les 2 analyses ne concordent pas.

3. Modèles hiérarchiques généraux

Nous allons examiner 2 méthodes actuellement explorées par les biométriciens. L'une est fondée sur les probabilités conditionnelles: c'est l'approche bayésienne. L'autre se place dans le cadre de procédures d'estimations plus classiques.

3.1. Modèle hiérarchique bayésien

3.1.1. Formulation et interprétation

GATSONIS *et al.* **[30]** étudient, aux Etats-Unis, les différences inter-états (variabilité géographique) des taux d'utilisation de l'angiographie coronaire après infarctus aigu du myocarde chez des malades de plus de 65 ans. L'unité épidémiologique est l'état (51 unités). Deux covariables sont utilisées pour caractériser chaque patient: l'âge et le sexe. Environ 220.000 cas sont retenus, extraits d'un fichier rassemblant les personnes ayant été atteintes d'un infarctus aigu du myocarde en 1987.

L'objectif de l'analyse est d'expliquer la variabilité de l'utilisation de l'angiographie parmi les malades d'un même état, et entre-états. Les auteurs veulent aller plus loin que la mise en évidence de l'hétérogénéité des taux: ils souhaitent rendre compte, dans la formulation du modèle, de la structure hiérarchique des données: échelle de l'individu et échelle de l'état.

La variable à expliquer est dichotomique: mise en oeuvre ou pas de l'angiographie après un infarctus aigu du myocarde. La régression logistique est utilisée pour modéliser la variation intra-état, c'est-à-dire à l'échelle de l'individu. Il est postulé que les variations inter-états sont aléatoires. Deux échelles sont définies:

Echelle 1: variabilité intra-état (variabilité des individus d'un même état)

Soit $p_{i,j}$ la probabilité du patient **j** de l'état **i** de subir une angiographie. Les auteurs supposent que $p_{i,j}$ suit une loi logistique spécifique de l'état **i**:

$$logit\ p_{i,j} = \beta_{i,0} + \beta_{i,1} X_1 + \beta_{i,2} X_2 + \beta_{i,3} + \ldots + \beta_{i,k} X_k$$

où les X sont les covariables caractérisant les patients (âge, sexe) et leurs éventuels termes d'interaction. Il y a donc 51 modèles logistiques de ce type (1 par état).

Echelle 2: variabilité inter-état

Le modèle de la variabilité inter-état consiste à poser que les 51 vecteurs de $\mathbb{R}^{k+1}$:

$$\beta_i = (\beta_{i,0}, \beta_{i,1}, \ldots \beta_{i,k})',$$

sont chacun des tirages aléatoires indépendants, provenant d'une distribution normale multivariée de vecteur moyenne à **(k + 1)** composantes:

$$\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_k)',$$

et de matrice de dispersion **T** à **(k + 1)** lignes et **(k + 1)** colonnes.

Les auteurs supposent qu'aucune connaissance préalable ne fournit d'information sur les hyperparamètres de population α et **T** qui sont inconnus et que l'on souhaite estimer.

Chaque composante du vecteur moyenne α (inconnu) représente la moyenne nationale de chaque coefficient correspondant du modèle de régression logistique de chaque état. Ainsi, par exemple:

$$\alpha_2 = \frac{\sum_{i=1}^{51} \beta_{i,2}}{51}$$

Si chaque covariable est centrée sur sa moyenne nationale, la composante α_0 représente la moyenne sur les 51 états du logarithme de l'odds ratio (log-odds) de l'angiographie pour le "patient moyen" hypothétique, c'est-à-dire dont chaque covariable aurait pour valeur la moyenne nationale correspondante.

T est la matrice des variances et covariances des coefficients β_i: le premier élément de la diagonale de **T** est égal à la variance du terme constant $\beta_{i,0}$, $Var(\beta_{i,0})$, de la série des 51 régressions logistiques. Le second élément correspond à $Var(\beta_{i,1})$, etc.

Le modèle peut être utilisé pour estimer la distribution de la probabilité de subir une angiographie dans chaque strate de malades, définie par un vecteur c_j de

covariables ***X*** fixes (âge et sexe données). Les log-odds de la probabilité de subir une angiographie pour des malades d'âge et de sexe donnés d'un état ***i*** sont calculés par les produits $\mathbf{c_j}' \boldsymbol{\beta_i}$. En conséquence des suppositions du modèle de l'échelle 2, ces log-odds ont une distribution normale de moyenne $\mathbf{c_j}' \boldsymbol{\alpha}$ et de variance $\mathbf{c_j}' \mathbf{T}\, \mathbf{c_j}$.

3.1.2. Estimation des paramètres et vérifications

L'estimation des paramètres de l'échelle 1 est obtenue en ajustant les modèles de régression logistique aux données de chaque état, par la méthode du maximum de vraisemblance. Les covariables sont centrées sur leurs moyennes nationales respectives.

Pour l'estimation des paramètres de l'échelle 2 ($\boldsymbol{\alpha}$ et **T**), les auteurs se placent dans un cadre d'inférence bayésien, dont on peut, par exemple, trouver un exposé dans un article de BERNARDINELLI et MONTOMOLI **[31]**. Ils comparent 2 techniques:

☞ **une analyse bayésienne utilisant l'échantillonneur de GIBBS.** Cette analyse comporte 2 étapes:

- **1. Simulations des distributions conditionnelles** des paramètres $\boldsymbol{\alpha}$, ***T*** et $\boldsymbol{\beta_i}$, avec ***i = 1,... 51***. Les simulations sont effectuées par un échantillonneur de GIBBS. Cette méthode de simulation stochastique s'apparente à la technique de Monté-Carlo. Elle consiste à construire une chaîne de MARKOV dont la distribution à l'équilibre est celle que l'on veut simuler. Cela permet de générer, par itérations successives, les réalisations d'une distribution multidimensionnelle.

- **2. Les moyennes d'échantillon** des valeurs simulées constituent les estimations de $\boldsymbol{\alpha}$ et ***T***. Ces valeurs sont utilisées pour établir des estimations du log odds de la probabilité de subir une angiographie pour les strates de malades définies par le sexe et l'âge.

Selon GATSONIS *et al.*, cette technique donne de bons résultats mais nécessite des calculs intensifs, que l'on ne pourrait pas effectuer en routine.

☞ **une approximation bayésienne empirique, utilisant un modèle normal des coefficients de régression logistique estimés à l'échelle 1.**

- **1. On estime les coefficients inconnus** $\boldsymbol{\beta_i}$ pour chacun des 51 états grâce aux modèles de régression logistique. Les tailles d'échantillon de chaque état étant grandes conditionnellement aux $\boldsymbol{\beta_i}$, le vecteur des coefficients estimés de chaque état suit une loi normale à ***(k + 1)*** dimensions, de moyenne le vecteur des coefficients $\boldsymbol{\beta_i}$, et de matrice de variance/covariance $\boldsymbol{V_i}$ (estimée d'après les coefficients des régressions logistiques).

- **2. D'après l'hypothèse de normalité faite pour l'échelle 2**, pour $\boldsymbol{\alpha}$ et ***T*** donnés, les estimations des β_i suivent des distributions indépendantes normales à ***(k+1)*** dimensions, de vecteur moyenne $\boldsymbol{\alpha}$ et de matrice de variance/covariance $\boldsymbol{T + V_i}$. Le centrage et la réduction des covariables des modèles logistiques de l'échelle 1 permettent d'obtenir des estimations des coefficients de régression (diagonale de $\boldsymbol{V_i}$) approximativement indépendantes. Cela

permet une estimation univariée, plus facile que l'estimation multivariée.

▸ **3. L'estimation de α et des éléments de la diagonale de T** sont effectuées par une méthode d'ajustement de la vraisemblance.

Les procédures classiques de vérification sont utilisées pour les modèles 1 (équations de régression logistique). Pour le modèle 2, les vérifications concernent surtout l'approximation bayésienne empirique, qui nécessite plus d'hypothèses que l'analyse bayésienne.

☞ **l'hypothèse d'indépendance des β_i** est vérifiée en examinant la matrice de corrélation T. Si elle contient des éléments négligeables en dehors de ceux de la diagonale, cela confirme l'absence de corrélation inter-états.

☞ **L'hypothèse de normalité des β_i** est vérifiée en traçant le graphe des valeurs centrées réduites des valeurs estimées des β_i en fonction des quantiles de distribution de la loi **N**(0, 1). Les auteurs testent également la normalité des combinaisons linéaires des composantes des estimations des β_i (toute combinaison linéaire de ces composantes doit être distribuée normalement si l'estimation de β_i suit une loi normale multivariée).

☞ **L'hypothèse de permutabilité** (absence de tendance). Ils font le graphe de ces mêmes valeurs centrées réduites en fonction de la taille des états.

L'approximation bayésienne empirique est plus simple en termes de calculs, mais les hypothèses doivent être soigneusement vérifiées. Elle est bien appropriée à l'exemple traité, où les effectifs sont élevés quels que soient les états et les strates de malades étudiés. Il n'en irait pas de même si des données étaient manquantes ou incertaines à cause de la faible taille des échantillons.

3.1.3. Bilan

Les suppositions sont contraignantes pour l'approximation bayésienne. L'indépendance des observations (hypothèse de permutabilité) est essentielle: cela limite l'intérêt en épidémiologie vétérinaire. En transposant dans l'exemple [Tchad], cela conduirait à admettre qu'il n'y a pas de hiérarchie structurant l'échelle du troupeau de concession. Cependant, cette difficulté peut être surmontée en élaborant des modèles plus complexes, développés pour tenir compte d'effets de gradients. Une revue des techniques actuellement disponibles est faite dans l'article de BERNADINELLI et MONTOMOLI **[31]**.

Par ailleurs, les calculs sont complexes, très longs et ne sont pas disponibles dans des logiciels commerciaux. BERNADINELLI et MONTOMOLI sont optimistes sur cet aspect, en tablant sur les progrès de l'informatique dans un proche avenir.

GATSONIS *et al.* n'ont pas cherché à identifier les facteurs de variation à l'échelle supérieure. Ils indiquent cependant, dans leur discussion, que leurs recherches actuelles vont vers la mise au point d'un modèle de régression de la moyenne des β_i, permettant ainsi la prise en compte de covariables caractéristiques des états.

Ce modèle offre des perspectives intéressantes, mais est encore du domaine de

la recherche biométrique. Il est difficilement applicable en épidémiologie vétérinaire dans les conditions actuelles.

3.2. Modèles de régression à plusieurs échelles

Ces modèles peuvent être vus comme des transpositions de modèles d'analyse hiérarchique de la variance (*nested analysis of variance*) décrits par SOKAL et ROHLF **[32]** (par exemple), pour des variables quantitatives. Le modèle d'analyse hiérarchique de la variance à 2 niveaux s'écrit:

$$Y_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk}$$

où:

Y_{ijk} k^{ème} observation du j^{ème} sous-groupe du i^{ème} groupe,
μ moyenne de la population,
A_i contribution aléatoire du i^{ème} groupe du niveau supérieur A de la hiérarchie,
B_{ij} contribution aléatoire du j^{ème} sous-groupe du i^{ème} groupe,
ϵ_{ijk} terme d'erreur associé au k^{ème} individu du j^{ème} sous-groupe du i^{ème} groupe

On suppose que A_i, B_{ij} et ϵ_{ijk} suivent des lois normales de moyennes nulles et de variances respectives σ_A^2, $(\sigma_{B \subset A})^2$ et σ^2.

La démarche exposée par SOKAL et ROHLF consiste à tester, dans un premier temps, la signification de l'effet B_{ij}, situé le plus bas dans la hiérarchie, sans tenir compte de l'effet A_i. Si aucun effet n'est mis en évidence, on arrête l'analyse. Dans le cas contraire, on poursuit en adoptant le point de vue inverse: on ne tient compte que de A_i. A l'issue de ces 2 étapes, on peut décomposer la variance totale en ses diverses composantes hiérarchiques.

3.2.1. Modèles linéaires mixtes

GOLDSTEIN **[33]** s'appuie sur l'exemple suivant: il considère un ensemble de données organisées en 3 échelles hiérarchiques: des écoles, des classes dans ces écoles et des enfants dans ces classes. Nous détaillons ce modèle, car cette structure est analogue à celle du système d'élevage caprin de l'enquête [Tchad]. Pour le j^{ème} enfant de la i^{ème} classe de la k^{ème} école, le modèle complet s'écrit:

$$Y_{kij} = \alpha^{\star}_{kij} + \beta^{\star}_{ki} + \gamma^{\star}_{k}$$

chaque échelle de la hiérarchie étant elle-même décrite par un modèle linéaire:

☞ **pour l'échelle de l'école:**

$$\gamma^{\star}_{k} = \gamma_0 + \gamma_1 w_{1,k} + \ldots + v_k = \sum_{l=0}^{q} \gamma_l w_{l,k} + v_k$$

où: v_k est un variable aléatoire telle que $E(v_k) = 0$ et $\mathrm{var}(v_k) = \sigma_v^2$,
et γ_l est le coefficient de la $l^{ème}$ covariable $w_{l,k}$ de l'école k

☞ **pour l'échelle de la classe:**

$$\beta^{\star}_{ki} = \beta_0 + \beta_{1,k} z_{1,ki} + \ldots + u_{ki} = \sum_{l=0}^{p} \beta_{l,k} z_{l,ki} + u_{ki}$$

où u_{ki} est un variable aléatoire telle que $E(u_{ki}) = 0$ et $\mathrm{var}(u_{ki}) = \sigma_u^2(k)$,
et $\beta_{l,k}$ est le coefficient de la $l^{ème}$ covariable $z_{l,ki}$ de la classe ki

☞ **pour l'échelle de l'enfant:**

$$\alpha^{\star}_{kij} = \alpha_0 + \alpha_{1,ki} x_{1,kij} + \ldots + e_{kij} = \sum_{l=0}^{r} \alpha_{l,ki} x_{l,kij} + e_{kij}$$

où e_{kij} est un variable aléatoire telle que $E(e_{kij}) = 0$ et $\mathrm{var}(e_{kij}) = \sigma^2(ki)$,
et $\alpha_{l,ki}$ est le coefficient de la $l^{ème}$ covariable $x_{l,kij}$ de l'enfant kij

Par substitution , le modèle complet s'écrit donc:

$$Y_{kij} = \alpha_0 + \beta_0 + \gamma_0 + \sum_{l=1}^{r} \alpha_{l,ki} x_{l,kij} + \sum_{l=1}^{p} \beta_{l,k} z_{l,ki} + \sum_{l=1}^{q} \gamma_l w_{l,k} + (v_k + u_{ki} + e_{kij})$$

soit en notation matricielle: $\boldsymbol{Y = X\beta + E}$

Ce modèle combine des effets fixes X et aléatoires E: c'est un modèle mixte. La variance de Y_{kij} ne dépend que des termes liés aux effets aléatoires. Si les covariances des effets aléatoires sont nulles, $var(Y_{kij}) = \sigma_v^2 + \sigma_u^2(k) + \sigma^2(ki)$. La variance globale de Y peut être décomposée en parties spécifiques de chaque échelle: école, classe et enfant. GOLDSTEIN indique comment exprimer ces composantes en fonction des variances σ_v^2 , $\sigma_u^2(k)$ et $\sigma^2(ki)$.

L'estimation des paramètres est fournie par un algorithme itératif des moindres carrés généralisés, que GOLDSTEIN a modifié ultérieurement **[34]** afin d'obtenir des estimateurs non biaisés (algorithme itératif des moindres carrés généralisés non biaisés restreints). Dans les 2 cas, les résultats sont identiques à ceux obtenus avec le maximum de vraisemblance (ou le maximum de vraisemblance restreinte) si les termes d'erreur suivent une loi normale multivariée.

On peut inclure des termes d'interaction entre variables de même niveau ou de niveaux différents, et des coefficients aléatoires. Le modèle s'exprime alors, en notation matricielle:

$$\boldsymbol{Y = X\beta + Ze},$$

Z et **e** sont les analogues aléatoires de **X** et **β**.

La formulation du modèle se rapproche de nos préoccupations. Toutefois, il ne permet pas de traiter les variables dichotomiques. Pour cela, il faut envisager des modèles hiérarchiques non linéaires mixtes.

3.2.2. Modèles non linéaires mixtes

3.2.2.1. Méthode

Le modèle hiérarchique non linéaire à effets mixtes peut être écrit comme la somme d'une composante non linéaire et d'une composante linéaire **[35]**:

$$y = f(X_1 \beta + Z_u u) + X_2 \gamma + Z_e e$$

f est une fonction non linéaire. β et γ sont des vecteurs de coefficients fixes, de matrices de réalisation correspondantes X_1 et X_2. e, u sont des ensembles de variables aléatoires d'espérances nulles et de matrices de réalisation correspondantes Z_e et Z_u. Les composantes de ces éléments sont les suivantes:
$\beta = (\beta_0, ..., \beta_l)^T$, $\gamma = (\gamma_0, ..., \gamma_m)^T$, $u = (u_1, ..., u_p)^T$ $e = (e_1, ..., e_q)^T$,
$X_1 = (x_{11}, ..., x_{1l})$, $X_2 = (x_{21}, ..., x_{2m})$, $Z_u = (z_{u1}, ..., z_{up})$, $Z_e = (z_{e1}, ..., z_{eq})$.

☞ Habituellement, Z_u et Z_e sont des sous-ensembles de X_1 et X_2, mais ce n'est pas indispensable. Z_u et Z_e peuvent avoir des vecteurs communs.

☞ Les variables aléatoires peuvent être relatives à n'importe quel niveau: dans certaines applications, ***e*** contient les vecteurs aléatoires de l'échelle 1 et ***u*** contient les vecteurs aléatoires de toutes les échelles supérieures.

Une application de ce modèle peut nous intéresser: il s'agit du modèle hiérarchique log linéaire à 2 échelles, où la variable dépendante est qualitative. Les composantes du vecteur réponse sont des proportions π_{hij}: une pour chaque cellule **i** du tableau de contingence à plusieurs entrées défini dans chaque unité **j** de l'échelle 2 (**h** = 2).

GOLDSTEIN écrit le modèle suivant pour la proportion moyenne de la $\mathbf{h}^{\text{ème}}$ unité de niveau 1 située dans la $\mathbf{i}^{\text{ème}}$ cellule de la $\mathbf{j}^{\text{ème}}$ unité du niveau 2:

$$\log(\pi_{hij}) = \sum_{k=0}^{l} \beta_{jk} x_{hijk}$$

$$\sum_i \pi_{hij} = 1; \quad h = 1, ..., q; \quad i = 1, ..., m_j$$

Il y a un total de $(q * m_j)$ unités du niveau 1 dans la $j^{\text{ème}}$ unité du niveau 2.

π_{hij} est la réponse prévue par le modèle. La proportion observée lors de l'enquête est p_{hij}, où n_{ij} est la taille de la $i^{ème}$ cellule de la $j^{ème}$ unité de la $2^{\text{ème}}$ échelle. Le nombre de réponses positives observées $n_{ij} p_{hij}$ est distribuée conditionnellement à i et j, selon une loi multinomiale de moyenne $n_{ij} \pi_{hij}$.

Certains coefficients β_{jk} (voire tous, si cela se justifie) peuvent être envisagés comme des variables aléatoires à l'échelle 2: $\beta_{jk} = \beta_k + u_{jk}$ (k = 1,... l). Les termes u_{jk} sont des variables aléatoires de l'échelle 2, de moyennes nulles, de distributions continues et de matrices de covariance finies.

Les x_{hijk} sont d'une part des variables de l'échelle 1, transformées en variables factices (*dummy variables*) pour les besoins de l'analyse, et d'autre part des covariables

de l'échelle 2 mesurées à cette échelle.

L'estimation des paramètres se fait en 2 temps. La première étape consiste à linéariser *f*. Les mêmes procédures que dans le modèle hiérarchique linéaire sont ensuite utilisées.

3.2.2.2. Exemple: analyse du chômage en Ecosse

La variable à expliquer est la proportion des non-chômeurs, caractérisés par 2 variables: le sexe et le niveau de qualification professionnelle (qualifié/non qualifié). 122 zones géographiques sont échantillonnées. Le modèle hiérarchique comporte 2 échelles: l'individu et la zone géographique. La zone géographique n'est décrite par aucune covariable. GOLDSTEIN écrit le modèle suivant:

$$\pi_{ij} = \frac{e^{(\beta_{j0} + \beta_1 x_{ij1} + \beta_2 x_{ij2})}}{1 + e^{(\beta_{j0} + \beta_1 x_{ij1} + \beta_2 x_{ij2})}} \qquad (i = 1,..., 4)$$

- x_{ij1} est une variable factice pour le sexe,
- x_{ij2} est une variable factice pour le niveau de qualification,
- π_{ij} est la proportion attendue dans la i^{eme} cellule du tableau de contingence 2 x 2 dans la $j^{ème}$ zone,
- le terme β_{0j} est aléatoire: $\beta_{0j} = \beta_0 + u_j$, où β_0 est la moyenne (inconnue) des β_{0j} dans les **j** zones étudiées, et u_j est une variable aléatoire ~ **N (0, σ^2)**

L'estimation des paramètres du modèle est résumée dans le tableau ci-contre. Les calculs sont effectués par un logiciel spécifique mis au point par RASBASH, PROSSER et GOLDSTEIN **[36]**. Les 4 effets aléatoires de l'échelle 1 correspondent aux 4 catégories différentes d'individus. Ce sont respectivement (1) les hommes non qualifiés, (2) les hommes qualifiés, (3) les femmes non qualifiées et (4) les femmes qualifiées.

Paramètre	Estimation	Ecart-type
Fixes		
"Constante"	0,522	
Sexe	0,148	0,11
Qualification	1,003	0,11
Aléatoires Echelle 2 (zones)		
σ_u^2	0,225	0,08
Echelle 1 (individus)		
σ_{e1}^2	1,20	0,20
σ_{e2}^2	0,94	0,16
σ_{e3}^2	0,88	0,14
σ_{e4}^2	1,09	0,16

3.2.3. Bilan

Le modèle hiérarchique non linéaire paraît susceptible de répondre à la plupart des situations rencontrées en épidémiologie vétérinaire, et particulièrement dans les enquêtes écopathologiques effectuées dans des systèmes d'élevage complexes.

Cependant, l'exemple fourni par GOLDSTEIN est peu démonstratif: aucune covariable n'est envisagée à l'échelle supérieure, et surtout, il ne compare pas ses résultats avec ceux que l'on pourrait obtenir avec un modèle non hiérarchique.

Une telle comparaison est effectuée par ALBANDAR et GOLDSTEIN dans un article d'application au domaine dentaire **[37]**. Les auteurs confrontent un modèle linéaire non hiérarchique et un modèle linéaire hiérarchique. Malheureusement, la partie méthodologique est peu détaillée et ne permet pas de comprendre toutes les hypothèses ayant présidé à la conception du modèle statistique. De plus, le modèle linéaire simple qui sert de point de comparaison n'est pas explicité dans cet article: il est fait référence à un autre article dont quelques résultats sont rappelés. Toutefois, il est indiqué que le modèle hiérarchique fournit des résultats plus simples en termes de parcimonie en variables. La variance des coefficients est plus forte qu'avec le modèle non hiérarchique, ce qui diminue la signification statistique, et permet d'éliminer des variables qui avaient été conservées avec ce modèle.

4. Conclusion

Pour conclure cet exposé, nous voudrions citer 2 exemples dont les enseignements sont convergents.

1. **Une revue internationale** attire l'essentiel des articles de résultats d'enquêtes épidémiologiques en médecine vétérinaire: il s'agit de *Preventive Veterinary Research* (Elsevier). Depuis 3 ans, 164 articles ont été publiés, dont 9 correspondaient au cadre fixé en introduction:

   ☞ au moins 2 niveaux hiérarchiques d'organisation différents,

   ☞ analyse simultanée des covariables définies à chacun des niveaux, intervenant sur la même variable à expliquer,

   ☞ variable à expliquer de nature dichotomique.

Les méthodes statistiques employées dans ces 9 articles étaient la régression logistique avec effet aléatoire (5 cas), la régression logistique ordinaire (1 cas), l'analyse des interactions étiologiques (2 cas) et l'analyse discriminante barycentrique (1 cas).

Il est clair que les méthodes hiérarchiques exposées dans la dernière partie de ce rapport ne sont pas encore passées dans la pratique. La principale raison de cette situation semble être que les connaissances biométriques elles-mêmes, mais également leur vulgarisation auprès des non-spécialistes, sont encore insuffisantes. Aucune méthode ne s'est imposée, et les algorithmes permettant l'estimation des paramètres nécessitent

encore de longs calculs sur les gros systèmes informatiques (en particulier pour les techniques bayésiennes).

Ce dernier aspect est d'ailleurs rédhibitoire pour les épidémiologistes travaillant dans les milieux tropicaux défavorisés, qui ne disposent que de micro-ordinateurs.

2. **Le premier congrès international d'écopathologie** s'est tenu en octobre 1993 à Clermont-Ferrand, organisé par le Laboratoire d'écopathologie de l'INRA. Environ 200 participants de 25 nationalités y ont assisté. La plupart des pays d'Europe occidentale, les Etats-Unis et le Canada étaient représentés, mais également de nombreux pays en développement (Algérie, Brésil, Inde, Maroc, Sénégal, Tchad, Zimbabwé). La question de la hiérarchie a souvent été abordée, beaucoup plus dans les discussions que dans les communications elles-mêmes. Les auteurs faisaient remarquer le manque d'outils, tant sur le plan théorique que sur celui des méthodes programmées dans les grands logiciels du commerce.

Les chercheurs souhaiteraient donc pouvoir analyser simultanément des covariables relevant de différents niveaux d'organisation d'un système hiérarchique, mais ils sont confrontés à un obstacle majeur de méthodologie statistique.

Dans le travail d'application de ce mémoire bibliographique, nous souhaiterions étudier la possibilité de surmonter ce problème. Dans un premier temps, nous voudrions détailler soigneusement l'étape de l'élaboration du pré-modèle conceptuel d'analyse.

Nous envisagerions ensuite de mettre en oeuvre une des méthodes décrites dans la troisième partie: la régression non linéaire hiérarchique, et la comparer à une méthode de référence non hiérarchique très utilisée depuis quelques années: la régression logistique avec variance extra-binômiale. Nous essayons actuellement de nous procurer les logiciels qui nous permettraient d'effectuer les calculs pour ces 2 techniques.

Remerciements

Je remercie mon Directeur de Recherches et mes parrains scientifiques pour leur aide précieuse et amicale.

Je remercie Mademoiselle Françoise LESCOURRET, chercheur au Laboratoire d'Ecopathologie, pour sa disponibilité constante, ses conseils pertinents, sa patiente pédagogie et sa bonne humeur.

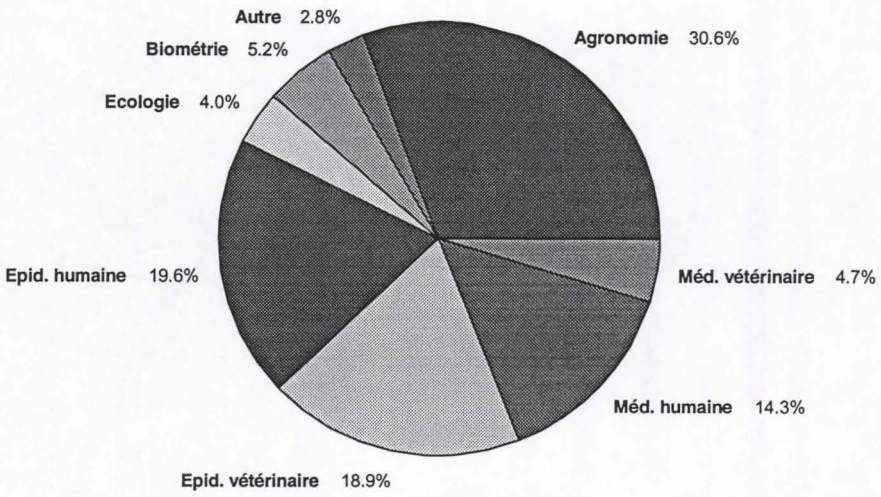
Je remercie le Directeur du CIRAD-EMVT et le Directeur du Laboratoire d'Ecopathologie de l'INRA, qui me permettent d'effectuer ce DEA dans les meilleures conditions possibles.

5. Annexe: bibliométrie

Pour les besoins de ce rapport, la recherche bibliographique a été effectuée selon 4 axes principaux.

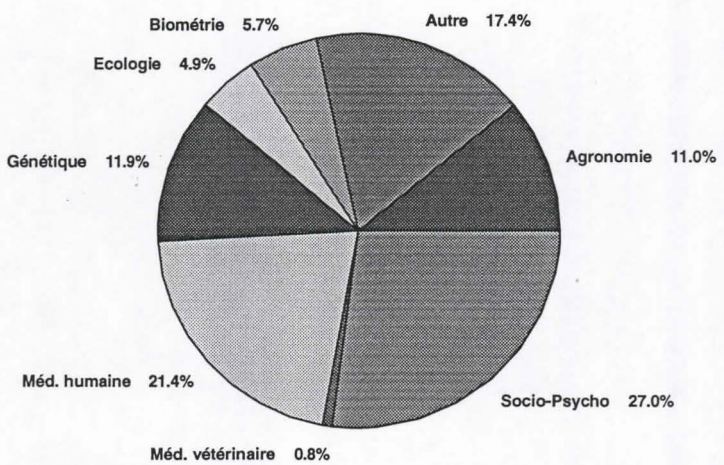
☞ Tout d'abord, 2 bases de données bibliographiques ont été consultées: CAB et MEDLINE, disponibles sur CD-ROM au centre INRA de Theix. 3 mots-clés ont servi à élaborer les requêtes: MODEL* (pour modèle ou modélisation...), HIERAR* (pour hiérarchie ou hiérarchique...) et EPIDEM* (pour épidémiologie ou épidémie ...).

Fig. 11. CAB, mots-clés EPID* et MODEL*
(n = 832)



Les tris croisés comportant les 3 mots-clés conduisent à des effectifs très restreints (7 références dans CAB). Ils ont donc été limités à 2 mots-clés: MODEL* et HIERAR* d'une part, et MODEL* et EPIDEM* d'autre part. Ces mots-clés ont été recherchés dans le titre et le résumé des publications. MODEL* et EPID* ont été recherchés uniquement dans CAB.

Fig. 12. CAB + MEDLINE, mots-clés HIERAR* et MODEL*
(n = 637)



Les résultats globaux sont présentés dans les 2 figures ci-dessus, où les articles ont été classés par grand domaine. Les résumés de ces 1.469 articles ont été examinés. 47 articles ont été jugés intéressants et ont été commandés, dont 8 ont été utilisés dans la rédaction de ce rapport.

☞ Les références citées dans chacun des 47 articles commandés ont été examinées, ce qui a conduit à commander une quarantaine de publications supplémentaires, dont 9 ont été utilisées pour ce texte.

Parmi les articles commandés, une forte proportion provient de revue de statistiques appliquées à la médecine.

☞ Tous les numéros de *Preventive Veterinary Medicine* ont été examinés (revue crée en 1985). Nous avons ainsi collecté une vingtaine d'articles, dont 6 ont été utilisés dans le rapport.

☞ Enfin, le Laboratoire d'écopathologie de l'INRA, où j'effectue mon stage de DEA, possède une riche collection bibliographique comprenant ses propres publications et la documentation rassemblée par les chercheurs. De plus, certaines références m'ont été recommandées par mes parrains pour ce DEA, et par les chercheurs avec lesquels je suis habituellement en contact. Une trentaine d'articles et ouvrages ont ainsi été réunis, dont 14 ont été utilisées.

Les 2 premières sources m'ont fourni les articles importants pour les bases méthodologiques, alors que les exemples d'applications ou des informations plus générales proviennent des 2 dernières. Si je devais refaire une nouvelle recherche bibliographique sur cette question, j'inclurais dans les mots-clés les termes MULTI-LEVEL et NON LINEAR. Toutefois, je ne pense pas que cela aurait fondamentalement modifié l'orientation du rapport, les articles supplémentaires se rapportant à ces mots-clés étant vraisemblablement très théoriques.

6. Bibliographie

1. INTERNATIONAL EPIDEMIOLOGY ASSOCIATION (LAST J.M., editor). *A dictionary of epidemiology. Second Edition.* New York, Oxford, Toronto, Oxford University Press, 141 p.

2. LANDAIS E., 1991. Ecopathologie et Systémique. *Etudes et Recherches sur les Systèmes Agraires*, (21):5-11.

3. GRAS R., BENOIT M., DEFFONTAINES J.P., DURU M., LAFARGE M., LANGLET A., OSTY P.L., 1989. *Le fait technique en agronomie.* PARIS, INRA/L'Harmattan, 183 p.

4. GANIERE J.-P., ANDRE-FONTAINE G., DROUIN P., FAYE B., MADEC F., ROSNER G., FOURICHON C., WANG B. et TILLON J.-P., 1991. L'écopathologie: une méthode d'approche de la santé en élevage. *INRA Prod. Anim.*, 4 (3):247-256.

5. MORLEY F.H.W., 1972. A systems approach to animal production: what is about? *Proc. Aust. Soc. Anim. Prod.* 9:1.

6. MARTIN S.W., MEEK A.H. and WILLEBERG P., 1987. *Veterinary Epidemiology. Principles and Methods.* Ames, Iowa State University Press, 343p.

7. HURD S.H., KANEENE J.B., 1993. The application of simulation models and systems analysis in epidemiology: a review. *Prev. Vet. Med.*,15: 81-89.

8. FAYE B., LEFEVRE P.C., LANCELOT R. et QUIRIN R., 1994. *Ecopathologie animale: la méthodologie. Applications en milieu tropical.* Versailles, INRA/CIRAD éd.

9. FRONTIER S., 1983. Introduction. *In: Stratégies d'échantillonnage en écologie*, Frontier S. (éd. scientif.), Paris, Masson, 1-11.

10. CORNFIELD J., 1978. Randomization by group: a formal analysis. *Am. J. Epidemiol.*, 108 (2): 100-102.

11. BENDIXEN P.H., 1989. The enigma of herd: a statistical problem or a question of study design ? *Prev. Vet. Med.*, 7: 69-71.

12. SCHERRER B., 1983. Chapitre 2. Techniques de sondage en écologie. *in Stratégies d'échantillonnage en écologie.* Frontier S. (éd. scientif.), Paris, Masson, p. 63-162.

13. GIGON A., 1987. A hierarchic approach in causal ecosystem analysis: the calcifuge-calcicole problem in Alpine grasslands. *In: Ecological studies, Vol. 61,* Berlin, SCHULZE and ZWÖLFER (ed.): 229-242.

14. JENICEK M. et CLEROUX F., 1982. Lexique anglais-français des termes utilisés en épidémiologie, *In: Epidémiologie: principes, techniques, applications.* Paris, Maloine S.A.: 447-450.

15. WRIGHT S., 1921. Correlation and causation. *J. Agric. Research*, 20: 557-585.

16. GOLDSMITH J.R., BERGLUND K., 1974. Epidemiological approach to multiple factor interactions in pulmonary disease: the potential usefulness of path analysis. *Annals New York Academy of Science*, 221: 361-375.

17. GOLDSMITH J.R., 1977. Paths of association in epidemiological analysis: application to health effects of environmental exposures. *Int. J. Epidemiol.*, 6 (4): 391-399.

18. LAFI S.Q. and KANEENE J.B., 1992. Epidemiological and economic study of the repeat breeder syndrome in Michigan dairy cattle. I. Epidemiological modeling. *Prev. Vet. Med.*, 14: 87-98.

19. BANGUERO H. Socio-economic factors associated with malaria in Colombia. *Soc. Sci. Med.*, 10: 1099-1104.

20. ALWIN D.F., HAUSER R.M., 1975. The decomposition of effects in path analysis. *Am. Sociol. Rev.*, 40: 37-47.

21. DONNER A., 1985. A regression approach to the analysis of data arising from cluster randomization. *Int. J. Epidemiol.*, 14 (2): 322-326.

22. HOSMER D.W., LEMESHOW S., 1989. *Applied logistic regression.* New York, Wiley ed., 307 p.

23. BOUYER J., 1991. La régression logistique en épidémiologie. Partie I. *Rev. Epidém. et Santé Publ.*, 39: 79-87.

24. BOUYER J., 1991. La régression logistique en épidémiologie. Partie II. *Rev. Epidém. et Santé Publ.*, 39: 183-196.

25. CURTIS C.R., ERB H.N., SNIFFEN C.J., SMITH D., KRONFELD D., 1985. Path analysis of dry period nutrition, post-partum metabolic and reproductive disorders, and mastitis in Holstein cows. *J. Dairy Sci.*, 68: 2347-2360.

26. CURTIS C.R., SCARLETT J.M., ERB H. and WHITE M.E., 1988. Path model of individual-calf risk factors for calfhood morbidity and mortality in New York Holstein herds. *Prev. Vet. Med.*, 6: 43-62.

27. CURTIS C.R., MAURITSEN R.H., SALMAN M.D. and ERB H.N. The enigma of herd: a comparison of different models to account for group effects in multiple logistic regression analysis. *Acta Veterinaria Scandinavica*, 84: 459-461.

28. WILLIAMS D.A., 1982. Extra-binomial variation in logistic linear models. *Appl. Statist.*, 31 (2): 144-148.

29. CURTIS C.R., ERB H.N., SCARLETT J.M. and WHITE M.E., 1993. Path model of herd-level risk factors for calfhood morbidity and mortality in New York Holstein herds. *Prev. Vet. Med.*, 16: 223-237.

30. GATSONIS C., NORMAND S.L., LIU C. and MORRIS C., 1993. Geographic variation of procedure utilization: a hierarchical model approach. *Medical Care*, 31 (5): YS54-YS59 (supplément).

31. BERNADINELLI L., MONTOMOLI C., 1992. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11: 983-1007.

32. SOKAL R.R, ROHLF R.J., 1981. Chapter 10. Nested analysis of variance. *In: Biometry. The principles and practice of statistics in biological research.* New York, W.H. Freeman and company, 2ème édition, p. 271-320.

33. GOLDSTEIN H., 1986. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73 (1): 43-56.

34. GOLDSTEIN H., 1989. Restricted unbiased iterative generalized least-squares estimation. *Biometrika*, 76 (3): 622-623.

35. GOLDSTEIN H., 1991. Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78 (1): 45-51.

36. RASBASH J., PROSSER R., GOLDSTEIN H., 1989. ML2 and ML3: software for two-level and three-level analysis. Users guide. London: Institute of Education.

37. ALBANDAR J., GOLDSTEIN H., 1992. Multi-level statistical models in studies of periodontal diseases. *J. Periodontol.*, 63: 690-695.